Ordinal-Contextual Dissimilarity for Analysis of Heros in Tragedies

Bahman Afsari
Dept. of Oncology, Johns Hopkins University
550 N. Broadway st., Baltimore, MD, 21205, USA
bahman@jhu.edu

Katayun Mazdapour
Inst. for Humanities Cultural Studies
63 Kurdistan Highway, Tehran, Iran

Bruno Jedynak
Dept. of Appl. Math. and Stats., Johns Hopkins University
3400 N. Charles st., Baltimore, MD, 21218, USA

Abstract

One [1] of the goals of a literary analysis of the tragedies is to identify the relationships among the heroes. We aim to provide computational tools which can help in this task. Ideally, this analysis would be based on the raw text. Instead, we rely on tables extracted by researchers from the text which summarize objectively the qualities of each hero from three famous classic stories in the available sources: Trojan War from the Homer's Iliad, Julius Caesar from the Shakespeare's play, and knights in the Ferdowsi's Shahnameh, a Persian epic story. We develop a statistical analysis for identifying clusters (groups) among a group of characters in a story starting from such tables with ordinal data. We define the Ordinal-Contextual (OC) similarity (or equivalently the OC dissimilarity) as a linguistically meaningful and context-based similarity (or dissimilarity) for ordinal data. Using classical multidimensional scaling, we represent the heroes in 2-D space and cluster the heroes using the affinity propagation technique. Overall, this analysis turns out to confirm and add quantitative validity to the hypothesis that the rival heroes in these classic tragedies who constitute the core of the tragedy are similar and distinguishable from the other heroes.

Introduction

This paper addresses a problem in the area of computational literary analysis. Literary analysis focuses on how plot, structure, character, and many other techniques are used by the author of a literary work to create meaning. The goal of a computational literary analysis is to broaden and deepen the understanding of a literary work using computational methods. The field of computational literary analysis is fairly new. For example, the first workshop on computational linguistics for literature took place in 2012 (NAACL in Montreal Canada). We propose an innovative computational method for the literary analysis of stories.

We focus on the relationships between the characters in a story. More specifically, we develop statistical tools for the literary analysis of the characters in the three famous classic tragedies: Trojan War (from Greek mythology) by Homer, Julius Caesar by Shakespeare and the Shahnameh, the Book of Kings (from Persian mythology) by Ferdowsi. The main technical contribution is our discussion of similarity (or equivalently dissimilarity) functions which would be relevant to the task of literary analysis as providing efficient tools for clustering and feature selection. In particular, we present a novel similarity function for ordinal data dependent on the context which is applied to cluster the heroes in the story and we found interesting patterns. Specifically, in all tragedies, the characters whose rivalry constitutes the core of the tragedy happen to be classified by automatic clustering in one small group. Hence, the quantitative analysis suggests that the main rival characters may represent similar characteristic to each other and very different from the rest of the characters in many tragedies.

[1] This work is dedicated to the late Master Valliollah Torabi and his passion for and contributions to saving Naghali, the traditional one-person show for the Shahnameh. We would want to gratefully thank R. Sarhangi for all his comments, help, and guidance. We would also like to thank B. N. Araabi and H. Hajimiri for their initial work and contributions to this work.

Ideally, all of our analysis would be automatic. However, this is a far reaching goal. In this work, we used summary tables provided by experts from reading the books. We present the data in the form of Tables 1, 2, and 3, in section "Data." In "Ordinal Contextual Analysis" section, we define *Ordinal-Contextual* (OC) similarity: a meaningful, intuitive and context-based similarity among characters in stories. By context-based, we mean not only does it depend on the two heroes of interest, but also it depends on the rest of the heroes. We find this in harmony with the nature of the human mind which compares people in the context of the rest of the people, not in an abstract and isolated world. In a more precise description, the OC similarity for two heroes measures how coherent two samples' features behave to the rest of the heroes. By ordinal, we refer to the fact that features are ordinal, i.e. there exists a sense of ranking in the value of features while we cannot assign the real numbers to the features representing their magnitude. We use the similarity for clustering the characters and utilize the induced OC dissimilarity to visualize the heroes relationships in a Euclidian two-dimensional (2D) space. In the "Discussion" section, we interpret the patterns found by the OC similarity analysis.

Data

To generate our data, we selected three classic stories which mainly are construed as tragedies: Trojan war from Iliad by Homer (Greek Mythology) [5], Julius Caesar [6] by William Shakespeare, and Knights in the Shahnameh by Ferdowsi (from Persian Mythology) [3]. As mentioned before, in an ideal setting, our analysis would be started from an automated text mining. However, this is not feasible at this moment. Instead, we rely on information extracted by experts who have read the texts and have mined mainly the events or characteristics of the heroes as objectively as possible. To do so, for Trojan war and Julius Caesar, we rely on a website which is well-known for summarizing the characters of classic stories called SparkNotes (http://www.sparknotes.com/). For the Shanameh, we relied on the data extracted by K. Mazdapour [3]. Moreover, in order to make the analysis under control and more meaningful, we only summarized main characters in the stories, i.e., characters involved in the war and power challenge, i.e. warriors, commanders and kings of both sides in Trojan War. For Julius Caesar, we mainly focused on the politician involved in the assassination of Caesar. For the Shahnameh, we only focused on the knights with repeated stories.

These tables describe important heroes in each story, one per row, with multiple ordinal features, one per column. Because of the space concerns, the columns and rows in the Shahnameh table represent the heroes and features respectively. Each feature represents a fact regarding each hero. For example, in the Shahnameh, table 3, the "Born by cesarean" row beneath the knight Rostam has a "Yes" because there is a story in the Shahnameh regarding his birth by cesarean, while the other knights, whose births were either not by cesarean or not described, have "No"s under their names in that row. As another example, "Seven labors" refers to an event narrated for both Rostam and Esfandiar, so these knights have "Yes" in that row under each of their names. In some cases, where the feature can be yes or no, there is still another label called "average" which means the poet narrated the story without much emphasis. The fact that features are ordinal numbers was considered in the definition of the OC similarity in "Ordinal Context Analysis" section.

Iliad is the first of two major poems written by ancient Greek poet, Homer. It is one of most influential work in the literature of western countries. Iliad describes the last part of the Trojan war which happened between Trojans under Priam and Achenes (i.e. Greeks) under Agamemnon. Achaeans are a coalition of cities which have been united against Trojans because a Trojan prince, Paris, has seduced and eloped with Helen, the wife of the Sparta King. The biggest tragedy in this story is the battle between Achilles, a legendary commander of Achaeans, and Hector, a legendary commander and the Crown Prince of Trojan. Achilles slain Hector who has a close personality to himself. Table 1 represents the data from Trojan war.

Julius Caesar is a tragedy written by William Shakespeare, arguably the most influential writer in English language. The play narrates the story of the conspiracy and assassination of Julius Caesar, a successful

Table 1 : *Description of the heroes in Iliad using 8 characteristics (or features).*

Side in War	Heroes	Royalty	Military Ability	Military Rank	Human or non-Human	Wisdom	Self-confidence	Orator
Achaeans	Agamemnon	King	No	Leader	Human	Reckless	Arrogant	No
	Achilles	No	Superb	Top-commander	Half-Human	Reckless	Proud	No
	Odysseus	King	Warrior	Commander	Almost	Wise	Neutral	Yes
	Great Ajax	No	Powerful	Commander	Almost	Neutral	Neutral	No
	Menelaus	King	No	No	Human	Neutral	Humble	No
	Diomedes	King	Warrior	Commander	Human	Brave	Neutral	No
	Little Ajax	No	Warrior	Commander	Human	Neutral	Neutral	No
	Patroclus	No	Warrior	No	Human	Wise	Neutral	No
	Nestor	King	Warrior	Commander	Human	Wise	Neutral	Yes
Trojans	Hector	Crown Prince	Superb	Top-Commander	Human	Neutral	Proud	No
	Aeneas	No	Powerful	Warrior	Half-Human	Neutral	Neutral	No
	Paris	Prince	Warrior	No	Human	Reckless	Neutral	No
	Priam	King	No	Leader	Human	Wise	Humble	No
	Polydamas	No	Warrior	Commander	Human	Neutral	Neutral	No
	Agenor	No	Warrior	No	Half-Human	Neutral	Neutral	No
	Sarpedon	No	Warrior	No	Half-Human	Neutral	Neutral	No
	Glaucus	No	Powerful	No	Human	Neutral	Neutral	No
	Antenor	No	No	Commander	Human	Wise	Neutral	No
	Polydorus	Prince	Warrior	Commander	Human	Neutral	Neutral	No
	Pandarus	No	Warrior	Commander	Human	Neutral	Neutral	No

and charismatic general and senator of the Roman Republic. The main tragedy of the story is the decision which Brutus, a scrupulous politician and a close friend of Caesar must make: He firmly believes in the Republic and thinks that the behavior of Caesar may lead to a dictatorship which finishes the Republic. Finally, Brutus joins the conspirators for the assassination. Table 2 represents the extracted data.

Table 2 : *Description of the eight characters of the Julius Caesar play using 8 characteristics (or features).*

Heroes	Senator	General	Orator	Conspirator	Scrupulous	Loyal	deceiving	Triumvirs
Brutus	No	No	No	Yes	Yes	Yes	No	No
Julius Caesar	Yes	Yes	No	No	No	Yes	No	No
Antony	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Octavius	No	No	No	Yes	No	No	Yes	Yes
Lepidus	No	No	No	Yes	No	Yes	Yes	Yes
Cassius	No	Yes	No	Yes	No	No	Yes	No
Cicero	Yes	No	Yes	No	Yes	Yes	No	No
Decius	No	No	No	Yes	No	No	Yes	No

The Shahnameh is, inarguably, the most influential mythological book for the Persian literature. The Shahnameh contains the legendary history of Iran. Its main subject is the great wars between the Iranian Empire and two other empires, the Turanian and the Roman Empires. The main heroes of these wars are knights. In the course of the narration of these wars, historical characters grow to become legendary personages. The peak of the heroism is the middle parts of the Shahnameh which is usually known as the epic part. The parts before the epic part are more mythical and the parts after epic part are called historical part. The knights appearing in the epic part of the Shahnameh, as well as their stories, are very famous in Iranian

popular culture. K. Mazdapour has gathered information about them from the text, which has been summarized in Table 3. The last tragedy of the epic part of the Shahnameh is the story of fight between two mighty knight Rostam and Esfandiar [4].

Table 3 : *Description of the six knights of the Shahnameh using twelve characteristics (or features).*

Characteristics	Saam	Rostam	Sohrab	Giv	Esfandiar	Rashnvad
Appearance in fights	A little	Much	Very Much	Little	Average	little
Born by cesarean	No	Yes	No	No	No	No
Being a knight in childhood	No	Yes	Yes	No	No	No
Armed in a specific way	No	Yes	No	No	Average	No
Story of finding a mace	Average	Yes	No	No	No	No
Story of finding a horse	No	Yes	Average	No	No	No
Story of first achievement	No	Yes	No	No	Average	No
Seven labors	No	Yes	No	No	Yes	No
Story of conquering a castle	No	Yes	No	No	Average	No
Identified as a supreme heroes	Average	Yes	No	No	Average	No
Honored as the world's knight	Yes	Yes	No	No	Yes	No
Honored as the defender of the crown	Almost	Yes	No	Almost	Yes	Little

Ordinal Contextual Analysis

In this part, we develop new methods to extract information from Tables 1, 2, and 3. We will address the problem of clustering the knights as well as representing them in a low dimensional Euclidian space for visualization. Clustering refers to a methodology in statistical analysis which finds patterns in data through assigning each sample (i.e. hero in our paper) to a group, also called a class or a cluster. We argue that traditional statistical methods (such as PCA and k-mean) are not helpful in analyzing these tables because they use operations that are meaningless in the realm of linguistics. So, we define the OC similarity (or equivalently the OC dissimilarity) between heroes which has three properties: first, it is meaningful for ordinal data; secondly, it takes into account the context, the other heroes; thirdly, the dissimilarity has desirable geometric (known as metric) property. For the sake of space, the proof is not shown here. Based on the OC similarity, we use an exemplar clustering method presented by [2]. In exemplar clustering, each cluster is defined by a prototype which is one of the heroes. So, the prototype is a meaningful center in contrast to the "center of gravity" of several heroes, not a meaningful quantity which traditional clusterings may find. We then embed the heroes in 2D Euclidian space using classical multidimensional scaling.

One shortcoming of the traditional methods which makes them unintelligible for Tables 1, 2, and 3 goes back to the fact that the entries in the tables are ordinal and not cardinal. There is an obvious understanding of more or less or equal in these features; however, the authors are not sure how meaningful attributing cardinal numbers (i.e. real numbers) to the features will be. Another shortcoming is related to the popular understanding of the concept of "difference" between two heroes; this concept, which must be the base and the source for defining a meaningful dissimilarity between any two heroes, must be defined in the context of the other heroes. In other words, we consider that the human mind compares two people in the context of all people it knows. On these bases, we propose the OC dissimilarity, which can be applied both to ordinal numbers and take into account the context.

OC (Dis-)Similarity Definition

All the features in three data tables are qualitative and ordered. For example, the feature "Armed in a specific way" in Table 3 is rated as "No", "Average" or "Yes." One can compare any two knights in the Shahnameh

for this feature. There are three possibilities: The first one can have a lower, a greater or the same value as the second one. We call the outcome of this comparison, the "relation" between two heroes in that feature. For example, the relation between Saam and Rashnvad in the feature "Honored as the defender of the crown" is that Saam is greater than "Rashnvad." Using the concept of relation, we define the similarity with respect to a single feature: the OC similarity between two heroes of interest with respect to a given feature is the number of other knights whose relation to these two heroes are the same in respect to that feature. For example, for the feature "Armed in a specific way," Rostam has the label "Yes," Esfandiar has the label "Average" and the other 4 knights have the label "No." Rostam and Esfandiar similarity adds up to 4 because they are both more "Armed in a specific way" than any of the remaining knights. The knights other than Rostam and Esfandiar have also a similarity of 4 to each other while the similarity between Saam and Esfandiar is only 1. The notion of Othe C similarity can be extended to multiple features by defining the similarity between two heroes as the sum of their similarities over all m features. More formally, we define the comparison function for two ordinal quantities a and b:

$$\delta(a,b) \triangleq \begin{cases} 1 & a > b \\ 0 & a = b \\ -1 & a < b \end{cases}$$

We now define the OC similarity of two heroes with indices j and k in one feature with index i.

$$S_i(j,k) \triangleq \sum_{l \in N \setminus \{j,k\}} I\{\delta(x_{ij}, x_{il}) = \delta(x_{ik}, x_{il})\} \tag{1}$$

where x_{ij} represents the feature i of hero j (an ordinal number), $I(p) = 1$ when proposition p is true and $I(p) = 0$ otherwise and N is the set of indices of heroes, $N = \{1, \ldots, n\}$. By summing up the similarities for each feature, we define the similarity between two instances indexed j and k in N as $S(j,k) \triangleq \sum_{i=1}^{m} S_i(j,k)$. Subsequently, since many theories and methods have been built on dissimilarity rather than similarity measures, we can define the OC dissimilarity (induced by the OC similarity) as $D(j,k) \triangleq S(k,k) - S(j,k)$. We can show that the OC dissimilarity has a nice mathematical (known as metric) property. For the sake of space, we skipped the proof. For each of Tables 1, 2, and 3, we can form another table called inter-point dissimilarity matrix. The inter-point dissimilarity matrix contains the dissimilarity for all pairwise heroes. Table 4 contains the inter-point dissimilarity matrix for all pairs of knights in the Shahnameh. Although inter-point dissimilarity matrix is a useful tool, it is preferable to have more visual tool to present the matrix. To do so, we used a standard technique called classical multi-dimensional scaling (MDS) which represents each hero as a point in 2-D plot and arrange the coordination of the points, so that the points' relative Euclidean distance is approximately close to what was represented in the inter-point dissimilarity matrix. The 2-D visualizations of Tables 1, 2, and 3 are shown in figure 1a, figure 2a and figure 2b. The approximated dissimilarity for the Shahnameh knights (i.e. distances in figure 2b) are also represented in Table 4 within parentheses. The visualization figures also contain the clustering information of heroes, that is, heroes clustered into one group are marked by the same symbol. This information can reveal some unbiased grouping solely based on the summary Tables 1, 2, and 3. Note that some clustering does not fully match the visualization because the 2-D visualization only carry a certain percentage of the variance of the data(in Iliad 56%, in Julius Caesar 74% and in the Shahanemeh 78%). In the following section, we try to interpret this information.

Intuitively, the OC similarity implies that two heroes are similar whenever they are relatively equal in most of the features. In fact, the context, the other heroes, plays a role in calculating the OC similarity. The OC similarity takes the maximum number whenever two instances are relatively equal to the other heroes in all features. Equivalently, the OC dissimilarity takes zero if and only if it compares a hero with him/herself.

Heroes	Saam	Rostam	Sohrab	Giv	Esfandiar	Rashnvad
Saam	0(0)	39(38.2)	31(27.5)	21(15.2)	27(18)	22(16.9)
Rostam	39(38.2)	0(0)	43(44.2)	52(51.6)	27(27.7)	53(52.5)
Sohrab	31(27.5)	43(44.2)	0(0)	24(24.9)	41(40.6)	22(23.7)
Giv	21(15.2)	52(51.6)	24(24.9)	0(0)	33(33.1)	12(2.2)
Esfandiar	27(18.0)	27(27.7)	41(40.6)	33(33.1)	0(0)	34(34.9)
Rashnvad	22(16.9)	53(52.5)	22(23.7)	12(2.2)	34(34.9)	0(0)

Table 4 : *The OC dissimilarity matrix from the Shahnameh. The numbers in the parentheses "()" are obtained using the embedding algorithm in 2D.*

Since we are using ordinal numbers, the labels in each feature can be replaced by their ranks in the set of labels for that feature. Rank-based methods have long been proven to be very efficient in non parametric statistical analysis ([7]). More recently, they were successfully used in the prognosis of cancer ([1]). Indeed, these methods are very robust and can be used in problems with long vectors and a limited number of instances. Obviously, our problem is different in nature from the gene expression problems. There is a very strong assumption in the cancer prognosis problem that only few features are related to the class labels, while in studying the classic stories, all features are obviously affecting the result and all characters are interconnected.

Discussion: Rival and similar characters

The main outcome of our analysis are visualizations in figures 1 and 2. We interpret the patterns found by our automatic analysis based on the clustering information and visualization. The general pattern is that important (and usual rival) characters in these three classic stories form their own small group of heroes. Especially, in all three stories, the main challenging characters which form the core of the tragedies were clustered together. In other words, our analysis supports quantitatively the hypothesis that in many classic tragedies two similar heroes (i.e. protagonist and antagonist) challenge each other. In fact, this can be the element which brings the suspense to the reader's mind.

Figure 1a represents the visualization for the heroes in the Trojan War. As can be seen, one group (Hector, Agamemnon and Priam) accompanied by Achilles occupy half of the plain while the rest of the heroes are much closer together. In fact, we know that probably these four are the most important heroes in the war. We can say that this group form the important kings who lead the war and the war is based on their hostility. Nestor and Odysseus form the less important commandars/kings group. The group consisting of Achilles, Great Ajax, Aeneas, Agenor, Glaucus, and Sarpedon is the important warrior group. The other group can be identified as less influential characters in the war: Menelaus, Diomedes, Little Ajax, Patroclus, Polydamas, Antenor, Polydorus, Pandarus, and Paris. Note that we infer that these heroes have less influential in the development of the story of the war but some of them are very important, e.g. Paris's eloping with Helen was the initiation of the war while he is not very influential in the development of the story.

Many scholar sources state that Hector and Achilles remind the readers of each other to some extent [5]. Although they did not show up in our analysis in the same group, we should note that Hector has also a royal character which makes him different from Achilles. One interesting experiment is to treat Hector with no royal background. The visualization is shown in figure 1b. The identified pattern is also meaningful. First, Hector and Achilles form one group whose fight establishes the core of the tragedy. Also, Agamemnon and Priam with Menelaus form the leaders/kings of the war and the rest of the clustering remains the same as before.

Figure 2a visualizes heroes in Julius Caesar play. As can be seen the core of the tragedy (Caesar

and Brutus) happen to be grouped together with Cicero. An interpretation can be that Cicero, similar to Brutus, is a scrupulous politician who honestly defends Caesar, unlike Brutus who honestly participate in the assassination. Either of the choices led to a tragic ending for the scrupulous politician. We can interpret Antony's singleton group as he is the pragmatic politician who wants to usurp the power which has a huge influence in the story but he is not essential to the tragedy that Brutus, a close friend of Caesar, participates in the conspiracy against him. The rest of the heroes are classified as one group which may be interpreted that they are less important characters.

Figure 2b visualizes the knights in the Shanameh. Rostam and Esfandiar formed one group and the rest of the knights another group. It is well known that Rostam and Esfandiar are the most important knights in the Shahnameh and other knights play a minor role. Again, we can see the same pattern that the characters who form the core of the tragedy happen to form a group. In fact, according to [3], although the rest of the knights are powerful, the most important role they have is to promote Rostam and Esfandiar through losing fights to these two superior knights. The fact that Rostam and Esfandiar form a superior class of heros in the mind of Iranian people has been studied and discussed in many sources, e.g. in [4].

Conclusion

In this paper, we analyzed tables extracted from three classic tragedies. These tables consist of features for heroes in the stories. To study these heroes, we introduce a new similarity measure between heroes called the OC similarity. The OC similarity measures the harmony of two ordinal samples in respect to the rest of the samples. We use the "affinity propagation" method for clustering. In order to visualize the results, we define the OC dissimilarity induced by the OC similarity. We use classical multidimensional scaling to visualize the heroes in 2-D. Interestingly, the result of the clustering found that the heroes whose challenge constitutes the core of the tragedy form a group. This quantitative analysis suggests that core of many tragedies is a challenge between two similar heroes who also happen to be different from other characters. We believe that these results should motivate further statistical analysis of stories by the OC similarity. Another future work is to modify the OC similarity which can handle cultural contexts in the stories.

References

[1] B. Afsari, U. B. Neto, and D. Geman. Rank discriminants for predicting phenotypes from rna expression. *Annals of Applied Statistics*, 8(3):1469–1491, 2014.

[2] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, Feb 2007.

[3] K. Mazdapour. *Dagh Gol Sorkh- Chahardah goftar digar darbareye arastoo (The heartbreak of the rose flower and fourteen other articles about mythologies)*. Asatir, 2004. In Persian.

[4] T. Noldeke. *Iranian National Epic or Shahnameh*. Porcupine Press, 1979.

[5] Public. Illiad. SparkNotes, study guide website, Winter 2015. http://www.sparknotes.com/lit/iliad/.

[6] Public. Julius caesar. SparkNotes, study guide website, Winter 2015. http://www.sparknotes.com/shakespeare/juliuscaesar/.

[7] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, pages 80–83, 1945.

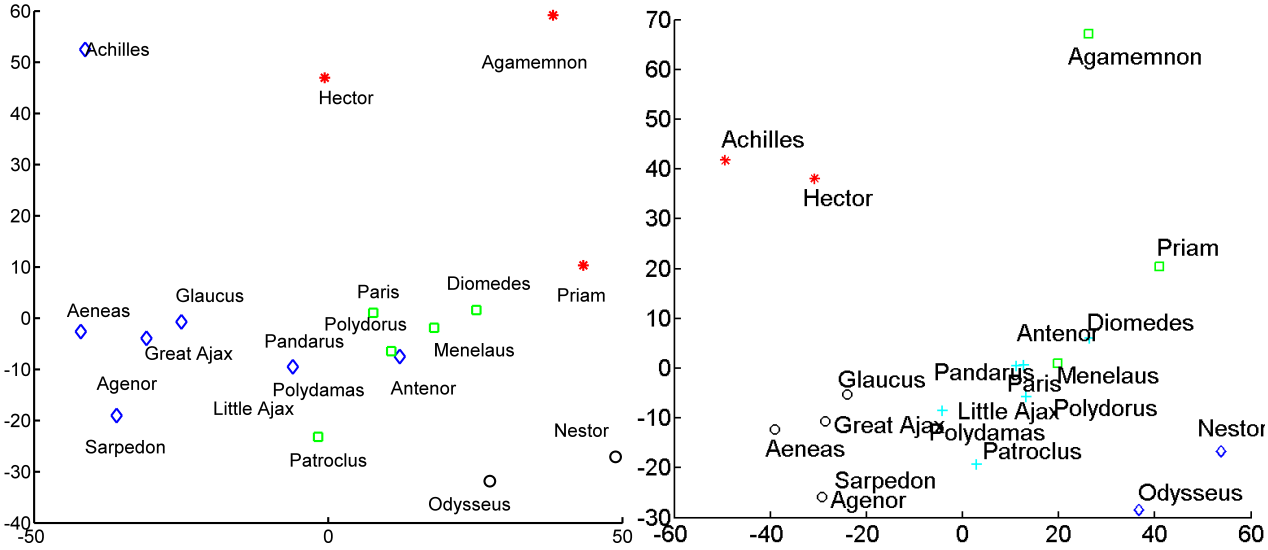

Figure 1 : ***(a)*** *2-D Visualization of heroes in the Trojan War. Hector, Agamemnon, and Menelaus can be construed as leaders/commandars of the war (marked by stars). Achilles and his group (marked by diamonds) can be construed as warriors. Nestor and Odysseus can be construed as minor commanders (marked by circles) and the rest are minor warriors.* ***(b)*** *The same visualization with exception of neglecting the Hector's royal background. Hector and Achilles form a group as the main heroes in the tragedy and Agamemnon, Priam and Menelaus form the leaders of the war.*

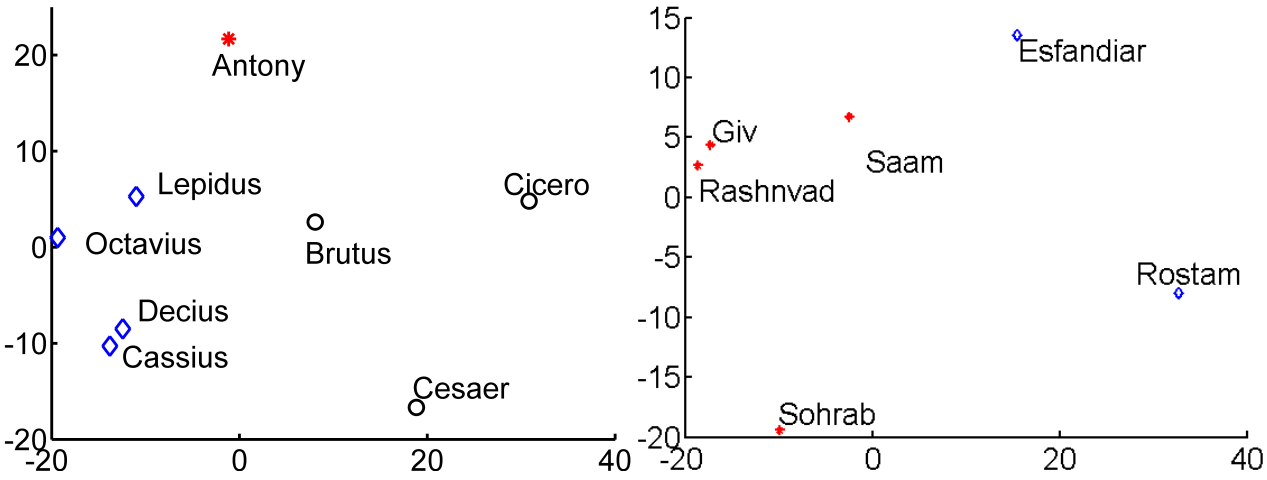

Figure 2 : ***(a)*** *Visualization for heroes in Julius Caesar play. Brutus, Caesar and Cicero form a group. The challenge between Brutus and Caesar form the core of the tragedy while Cicero may remind the reader of what would have happen to Brutus if he had not joined the conspiracy. Mark Antony is a singleton group presumably since he is an unscrupulous and pragmatic politician who usurps the power in any way.* ***(b)*** *Visualization for the Shahnameh heroes: Rostam and Esfandiar who constitute the core of the tragedy form a group.*